<https://www.kaggle.com/datasets/ankitkalauni/predict-the-data-scientists-salary-in-india>

For the dataset on predicting data scientists' salaries in India, here are some potential topics you can explore for your term project, focusing on link prediction, classification, and clustering:

### 1. \*\*Classification:\*\*

- \*\*Predict Salary Range\*\*: Build a classification model to predict whether a data scientist's salary falls into low, medium, or high ranges based on their experience, location, skills, and other features.

- \*\*Job Title Classification\*\*: If the dataset contains job titles, you could classify the roles into different types (e.g., junior, senior, lead) based on salary and experience.

- \*\*Skill-based Classification\*\*: Predict the likelihood of a candidate securing a specific salary bracket based on their skill set.

### 2. \*\*Clustering:\*\*

- \*\*Employee Segmentation\*\*: Use clustering algorithms to group data scientists into different categories based on salary, experience, and skills. This can help understand different "clusters" of employees based on factors such as skill levels and locations.

- \*\*Market Segmentation\*\*: Perform clustering to identify salary patterns across regions or industries. This can be useful in analyzing geographical or industry-based salary trends.

- \*\*Career Path Segmentation\*\*: Group data scientists based on their career trajectory—e.g., years of experience versus salary growth—to find common paths to high-paying roles.

### 3. \*\*Link Prediction:\*\*

- \*\*Skill-Salary Relationship\*\*: Treat each data point (skills, experience) as nodes, and predict links (relationships) between certain skills and salary outcomes. For example, predicting the likelihood of a certain combination of skills leading to a specific salary range.

- \*\*Location-Salary Trends\*\*: Predict the salary differences between regions and job locations based on company or city-specific trends using link prediction to explore relationships between location nodes and salary nodes.

Each of these topics could be explored in depth and would align with your current project constraints of classification, clustering, and link prediction. Let me know which one resonates with you, or if you'd like further elaboration!

Setting a \*\*performance target value\*\* for your data mining project can be done by referencing competitive benchmarks from Kaggle competitions or scholarly papers. These references ensure your target is grounded in existing research and reflects realistic expectations based on similar datasets and tasks. Here’s how you can approach it:

### 1. \*\*Kaggle Competitions\*\*

Kaggle competitions often provide leaderboards, which can help you gauge reasonable performance targets. You can find similar datasets related to salary prediction, classification, clustering, or link prediction. Look at the winning scores (like accuracy, F1 score, or AUC) and set your target relative to the top 10% or median performances, depending on your ambition.

\*\*Example steps\*\*:

- Search for competitions similar to "salary prediction" or "classification of professionals" on Kaggle.

- Note down the performance metrics (e.g., Accuracy: 85%, F1-score: 0.75).

- Set your target slightly below the top-tier performers but within range (e.g., if top teams achieve 90% accuracy, aim for 80-85%).

\*\*Competitions to reference\*\*:

- \*\*Predict Future Sales\*\*: Though it's about sales, the structure and techniques for predictive modeling can provide useful benchmarks for salary predictions.

- \*\*House Prices - Advanced Regression Techniques\*\*: This involves predicting continuous values (like salary), and the performance benchmarks can guide your targets for regression-based salary predictions.

### 2. \*\*Research Papers\*\*

Academic papers often report performance metrics for classification, clustering, and link prediction tasks. These papers provide a baseline and performance expectations using standard evaluation metrics like accuracy, precision, recall, or cluster purity.

\*\*How to find relevant papers\*\*:

- Search for papers on \*\*Google Scholar\*\* using keywords like "data scientist salary prediction," "classification in human resource analytics," or "link prediction for professional networks."

- Extract the reported performance values from papers that use similar datasets or tasks.

\*\*Key academic references\*\*:

- \*\*Salary Prediction Models\*\*: Papers focusing on salary prediction, such as “Using Machine Learning for Salary Prediction in HR Analytics” or “Predicting Salary from Resume Text and Job Descriptions.”

- \*\*Link Prediction\*\*: Research around "Link Prediction in Social Networks" can help in setting benchmarks for your link prediction tasks.

- \*\*Clustering for Employee Segmentation\*\*: Look for clustering-based HR analytics studies, such as “Clustering Techniques for Human Resource Data Analysis.”

### 3. \*\*Industry Benchmarks\*\*

You can also use salary benchmarking reports from the HR domain, which often use regression models or classification techniques. They may not provide explicit metrics like accuracy, but they help understand the variations in salary predictions across industries.

### Typical Performance Metrics to Consider:

- \*\*Classification\*\*: Accuracy, Precision, Recall, F1 Score, AUC-ROC.

- \*\*Clustering\*\*: Silhouette Score, Cluster Purity, Adjusted Rand Index.

- \*\*Link Prediction\*\*: AUC, Precision@k, Mean Reciprocal Rank (MRR).

### Steps to set a target:

1. \*\*Survey Kaggle competitions and papers\*\* on tasks similar to yours (e.g., classification, clustering).

2. \*\*Pick a reasonable benchmark\*\*: For instance, if the median accuracy in similar tasks is 85%, you can set a target of around 80-85%.

3. \*\*Justify your target\*\*: Cite the competition, leaderboard performance, or paper from which you derived your benchmark, and explain why it’s realistic based on your dataset and task.

If you'd like, I can help you search for specific competition results or papers to reference.